

Bootstrap

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution function of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples with replacement, of the observed dataset (and of equal size to the observed dataset).

Setting

- Dataset

$$\begin{pmatrix} X_1 & Y_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ X_N & Y_N \end{pmatrix}$$

- $M(X)$ prediction model
- Validation summary $Z(M, \varphi)$
- Estimator for $Z(M, \varphi)$: $Z(M, \hat{P})$ where \hat{P} is the empirical distribution using all the data
- Goal: Evaluating the level of uncertainty about $Z(M, \hat{P})$

Implementation

Step 1. Replace φ with an other distribution Q , where $Q = \hat{P}$

Step 2. Simulate iid data from Q

$$\begin{pmatrix} X_1^b & Y_1^b \\ \cdot & \cdot \\ \cdot & \cdot \\ X_N^b & Y_N^b \end{pmatrix}$$

where b denotes the data points generated at the b^{th} iteration, for $b = 1, \dots, B$.

Step 3. Monte Carlo Estimate of $Z(M, Q)$: $Z(M, \hat{q}_b)$ where \hat{q}_b is the empirical distribution of Q at the b^{th} iteration.

Step 4. Repeat steps 1, 2, and 3 B times and get the values $Z(M, \hat{q}_1), \dots, Z(M, \hat{q}_B)$.

Step 5. Use $Z(M, \hat{q}_1), \dots, Z(M, \hat{q}_B)$ to make inference on $Z(M, \varphi)$ (e.g. compute variance, standard deviation, etc.)

Step 6. Since Q has to be similar to φ , $Z(M, \hat{q}_b)$ has to be “similar” in distribution to $Z(M, \hat{P})$. Thus, the level of variability of $Z(M, \hat{q}_b)$ and $Z(M, \hat{P})$ has to be similar.

Example: $V(Z(M, \hat{q}_b)) = V(Z(M, \hat{P}))$

Example

Let's use the low birth weight data set from before.

```
library(MASS)
library(verification)
data(birthwt)
attach(birthwt)
str(birthwt)

## 'data.frame':  189 obs. of  10 variables:
## $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
## $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
## $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
## $ race : int  2 3 1 1 1 3 1 3 1 1 ...
## $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
## $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
## $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
## $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
## $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
## $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...

race = factor(race)
```

Suppose we are interested in a variance estimate for the Brier score we calculated in a previous example. The bootstrap can be implemented as follows:

```
B = 1000 # Number of iterations
brier.boot <- rep(NA, B) # Allocation vector

for(b in 1:B){
  # Resample
  idx = sample(1:nrow(birthwt), nrow(birthwt), replace=T)

  # Bootstrap sample
  x.bot = birthwt[idx,]

  # Fit the model
  mod.low = glm(low ~ lwt + race + age + ftv, data = x.bot, family=binomial(link=logit))
  pred = predict(mod.low, type='response')
  tt = verify(x.bot[,1], pred)

  # Evaluate the estimator
  brier.boot[b] = tt$bs
}
```

Let's look at the mean and standard deviation of our estimate, as well as a histogram of the distribution of values.

```
mean(brier.boot)
```

```
## [1] 0.1997093
```

```
sd(brier.boot)
```

```
## [1] 0.01338637
```

```
hist(brier.boot, freq = F, ylim = c(0,30), col = "grey")  
lines(density(brier.boot), lwd = 3, col = 'blue')
```

Histogram of brier.boot

